修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　　情報理工学研究科　　総合情報学専攻　博士前期課程		
氏　　名	小野　博紀	学籍番号	1230022
論 文 題 目	イベントの時系列分析による因果関係知識の獲得		

要　旨

因果関係に関する知識は、意思決定やリスク回避を行うとき、重要な知識となる。しかし、大規模な文書集合から人手で因果関係に関する知識を獲得するのはコストと時間がかかり、かつ多くの事象が絡み合っているため容易ではない。

そこで、文書集合から自動的に因果関係知識を獲得する研究が行われている。因果関係抽出に関する研究の多くは、因果関係にある事象対は文書内で共起しやすいという特性を利用しているため、文書内で共起しない事象対については考慮されていない。そのため、文書に明示的に書かれていないが人は常識として認識している因果関係や、新たな知識発見となる因果関係知識は抽出できない。そこで本研究では、明示的に書かれていない因果関係知識とともに、新たな知識発見となる因果関係知識を獲得する手法を提案する。

本研究においてイベントは「〈景気、上昇〉」のように名詞NPと動詞句VPのペアで表され、このペア〈NP, VP〉をイベント表現と呼ぶ。そして、提案する獲得手法では、新聞記事からイベント表現を抽出するところから始まる。次に、同じイベント表現を持つそれぞれのイベント表現集合に対して、話題ごとに分けるためのクラスタリングを行う。ここで得られたそれぞれのクラスタが、因果関係知識における1つの事象となる。そして、それらのクラスタ内に含まれるイベント表現を時系列順に並び替え、その時系列データに対して、イベントが話題となった時期を鮮明にするためにバースト検出を行う。最後に、これらの時系列データを対象にグレンジャー因果性検定を用いて因果性の有無を判断し、この検定で有意となったイベント対を因果関係知識として抽出する。

評価実験により、提案手法を用いることで妥当な因果関係とともに、既存手法では獲得しづらかった知識も獲得できることを確認できた。また、イベントの時系列データに対してバースト検出を行うことで、人では気づきにくい因果関係知識を獲得できることがわかった。

平成25年度総合情報学専攻博士前期課程修士論文

イベントの時系列分析による
因果関係知識の獲得

提出年月日：　平成26年1月30日
提出者：　学籍番号 1230022
小野 博紀
指導教員：　内海 彰 教授
尾内 理紀夫 教授

目 次

1　はじめに

因果関係に関する知識は，意思決定やリスク回避を行うとき，重要な知識となる．自然言語処理の分野においても，因果関係の知識は情報の整理や検索，質問応答システムなど，様々なアプリケーションにとって重要な知識源となる．

しかし，大規模な文書集合から人手で因果関係に関する知識を獲得するのはコストと時間がかかり，かつ多くの事象が絡み合っているため容易ではない．

そこで，文書集合から自動的に因果関係知識を獲得する研究が行われている．例えば,「に伴う」や「を理由に」といった手がかり表現を用いる手法 [1, 2] や，構文パターンを用いる手法 [3, 4]，動詞並列句に注目した手法 [5] などが提案されている．そして，これらの研究の多くは因果関係にある事象対は文書内で共起しやすいという特性を利用しているため，文書内で共起しない事象対については考慮されていない．

そのため,「金利が下がれば景気が上がる」といった，人は社会的常識として認識しているが明示的に文書に書かれにくい因果関係知識は獲得しづらい．また，明示的に文書内で述べられた事象から因果関係知識を獲得するため,「サブプライムローン問題によって経済悪化した」,「景気悪化によって内定取り消しが起きた」など，粒度が細かく，すでに既知な知識を獲得することが多い．

そこで本研究では，明示的に書かれていない因果関係知識とともに，新たな知識発見となる因果関係知識を抽出することを目的とする．この「明示的に書かれていない因果関係知識」には,「経営難の企業が増えると景気が悪化する」といった粒度の粗い知識も含めており，このような粒度の粗い知識は粒度の細かい知識よりも一般化された知識であるため，様々な言語アプリケーションへの応用が容易である．

提案手法は，新聞記事から「A が B になった」という単位の様々なイベントを収集後，各イベントが話題となっている時期を検出し，それらの時系列データ内で統計的に因果関係があると判断されたイベント対を因果関係知識として抽出する．つまり，イベント対に対して，文や文書内共起ではなく，時系列データに着目することで因果関係を抽出する．

本論文の以下では，2 章で関連研究について言及し，本研究の立ち位置について述べる．3 章では提案手法のアルゴリズムを示す．そして，4 章では提案手法を評価するための実験とその結果，その結果に対する考察について述べる．

2　関連研究

Giriju ら [1] や佐藤ら [2] は,「に伴う」や「cause」というような因果関係を表す手がかり表現を用いて，因果関係知識を獲得する手法を提案している．さらに坂地ら [3] や，Khoo ら [4] は因果関係を表す構文パターンを用いた知識獲得手法を提案している．

また，上記で挙げた研究とは違い，手がかり表現や構文パターンに依存しない手法も提案されている．Trisawa[5] は並列句が一つの文に存在し，並列句中の動詞が共通の目的語を持つ場合，因果関係が成立しやすいと仮定して，統計的に因果関係を抽出する手法を提案している．Chang[6] や山田ら [7] は因果関係にある単語ペアと構文構造を学習する手法を提案している．この手法は，並列句中の動詞が共通の目的語を持たない場合にも対処可能である．

さらに，単語単位ではなく，動詞や目的語を含んだイベント単位での因果関係抽出も行われている．Riaz ら [8] はイベントを表現する, 文で構成されたクラスタ間の因果関係を求める研究を行なっている．Do ら [9] は動詞が表現するイベントと，作成したルールを用いて生成した名詞化表現を抽出し，それぞれのイベント間に因果関係が存在するかどうかを相互情報量などを用いた計算式を用いることで判定，抽出している．

因果関係の抽出とともに，分類を行った研究もある．乾ら [10] は手がかり表現「ため」を用いて知識を抽出し，それらを 4 つの因果関係 (cause, effect, precond,means) に自動的に分類する手法を提案している．

これまでに挙げた関連研究は，因果関係にある事象対は文書内で共起しやすいという因果関係の出現特性 [11] を利用しているが，本研究ではその特性を利用しない．この特性を利用しないことによって，既知の因果関係知識だけでなく，新たな知識発見となる因果関係知識を獲得できることを期待できる．

因果関係の出現特性を利用せず，因果関係を抽出する試みもされている．Sun ら [12] は検索エンジンのクエリのログから因果関係を抽出する手法を提案している．この研究では，検索回数が急激に上昇した日時に何らかのイベントが発生したとみなし，そのイベント間の発生した時間に着目して因果関係の抽出を試みている．この研究の解析対象はクエリであるが，本研究では新聞記事を解析対象とする．

3　提案手法

3.1　概要

本研究において，イベントは名詞句 NP と動詞句 VP のペアで表され，このペアをイベント表現 $\langle NP, VP \rangle$ と呼ぶ．そして VP については,「上昇，下降，発生」のいづれかが入る．例えば,「景気が上がった」というイベントは「$\langle$ 景気，上昇 $\rangle$」というイベント表現で表される．

本手法は，大きく分けて以下の 4 つの段階で構成されている．

- イベント表現 $\langle NP, VP \rangle$ の収集
- イベント表現のクラスタリング
- 時系列順への並び替えとバースト検出
- 因果性の有無の判定

まず，新聞記事からイベント表現を抽出するところから始まる．次に，同じ $\langle NP, VP \rangle$ を持つそれぞれのイベント表現集合に対して，話題ごとに分けるためのクラスタリングを行う．このクラスタリングによって得られたそれぞれのクラスタが 1 つのイベントとなる．つまり，それぞれのクラスタが因果関係知識における 1 つの事象となる．そして，それらのクラスタ内に含まれるイベント表現を 1 週間を単位として時系列順に並び替え，その時系列データに対してバースト検出を行う．最後に，バースト検出によってイベントが話題となった時期が鮮明となったそれぞれの時系列データを対象に，グレンジャー因果性検定を用いて因果性の有無を判断する．この検定で有意となったイベント対には因果性があると判断し，因果関係知識として抽出する．

このように，本手法は既存の研究で用いられてきた因果関係である 2 つの事象は文書内で共起しやすいという前提を用いずに，イベントが発生した時期に対して分析を行うことで，因果関係知識を抽出する．本手法の全体図を図 1 に示す．

以下，それぞれの手順の詳細を述べる．

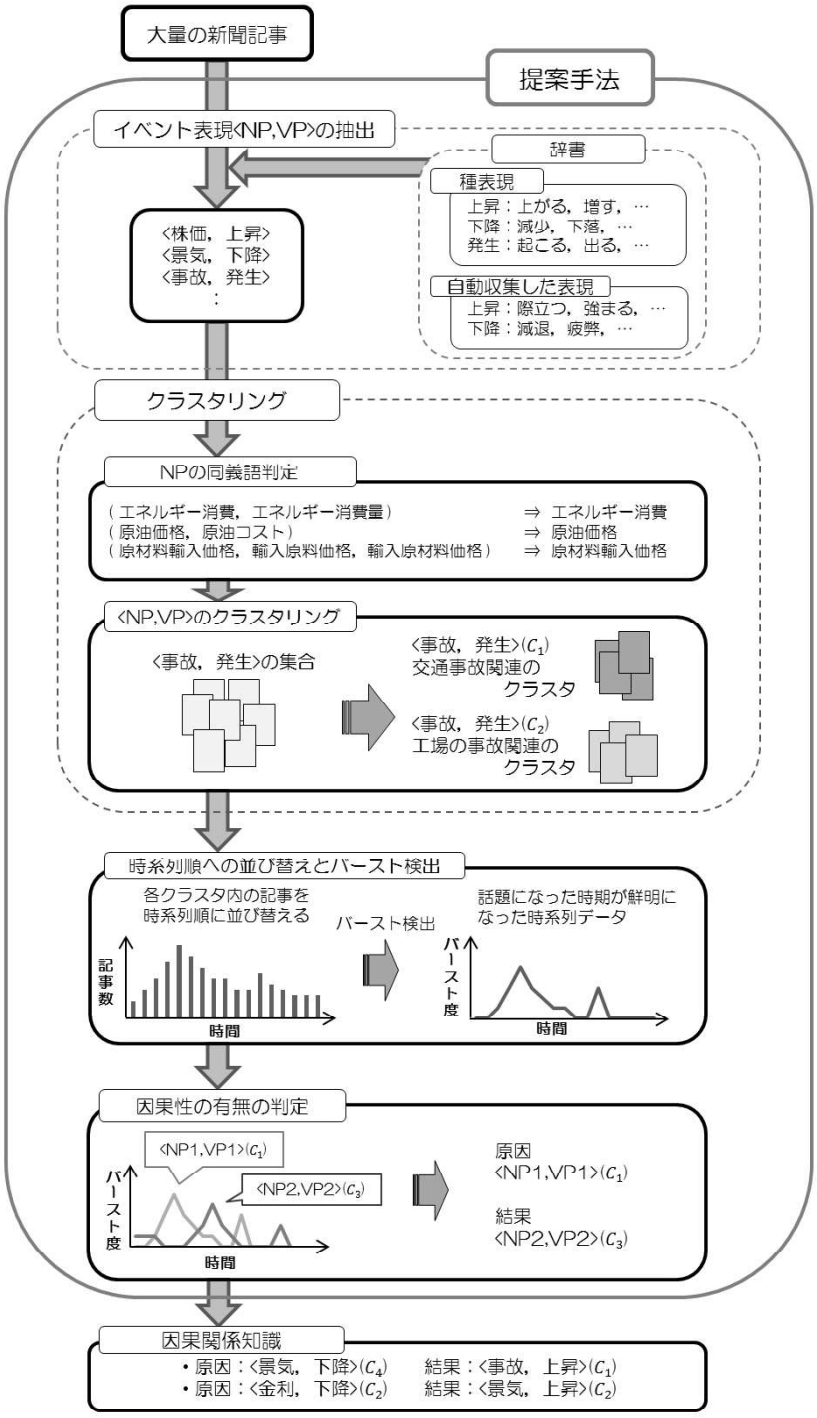
大量の新聞記事
提案手法
イベント表現<NP,VP>の抽出
辞書
種表現
上昇：上がる，増す，…
下降：減少，下落，…
発生：起こる，出る，…
自動収集した表現
上昇：際立つ，強まる，…
下降：減退，疲弊，…
<株価，上昇>
<景気，下降>
<事故，発生>
：
クラスタリング
NPの同義語判定
（エネルギー消費，エネルギー消費量） ⇒ エネルギー消費
（原油価格，原油コスト） ⇒ 原油価格
（原材料輸入価格，輸入原料価格，輸入原材料価格） ⇒ 原材料輸入価格
<NP,VP>のクラスタリング
<事故，発生>の集合
<事故，発生>(C_1)
交通事故関連の
クラスタ
<事故，発生>(C_2)
工場の事故関連の
クラスタ
時系列順への並び替えとバースト検出
各クラスタ内の記事を
時系列順に並び替える
バースト検出
話題になった時期が鮮明に
なった時系列データ
記事数
時間
バースト度
時間
因果性の有無の判定
<NP1,VP1>(C_1)
<NP2,VP2>(C_3)
バースト度
時間
原因
<NP1,VP1>(C_1)
結果
<NP2,VP2>(C_3)
因果関係知識
・原因：<景気，下降>(C_4) 結果：<事故，上昇>(C_1)
・原因：<金利，下降>(C_2) 結果：<景気，上昇>(C_2)

図 1: 提案手法の全体図

3.2 イベント表現 $\langle NP, VP \rangle$ の収集

格助詞「が，は」を含む文節が動詞，及びサ変接続の名詞に係っていた場合，この名詞句 NP と動詞句 VP のペアを 1 つのイベント表現 $\langle NP, VP \rangle$ とする．そして，VP が「上昇，下降，発生」に分類されたイベントを抽出対象とする．これは，因果関係知識を構成するイベントとして，何かが変化したイベント，または何かが起こったイベントが適切であると考えられるからである．例えば，$\langle$ 一年, 過ぎる $\rangle$ というイベントがあった場合，VP（過ぎる）は「上昇，下降，発生」のいずれにも分類されないため，抽出対象とならないが，$\langle$ 景気, 上がる $\rangle$ は VP（上がる）が上昇に分類されるため，抽出対象となる．例えば,「経常利益が前年同期比 50 %減に落ち込んだ。」という文があった場合，この文は図 2 のように係り受け解析される．そして，格助詞「が」を含む文節が動詞「落ち込む」に係っているため，$\langle$ 経常利益, 下降 $\rangle$ というイベント表現が抽出される．

上記の条件のみでイベント表現の抽出を行うと「景気が上がれば，～なる」といった仮定の意味で出現したイベント表現も抽出されてしまうため，格助詞「が，は」を含む文節の係り先の文節内に，助動詞「た」の仮定形，もしくは接続助詞「ば」が含まれていた場合，抽出対象から除いた．また，抽出した NP が「こと」であった場合,「こと」と，その文節に係る文節を結合して 1 つの NP とする．これは,「見ること」のような表現の場合，2 つの文節に別れてしまうからである．

イベント表現の抽出元となる文書として，新聞記事を用いる．これは，新聞記事はイベントが発生した時期に，そのイベントに関する記事が出現するという特性を持つためである．

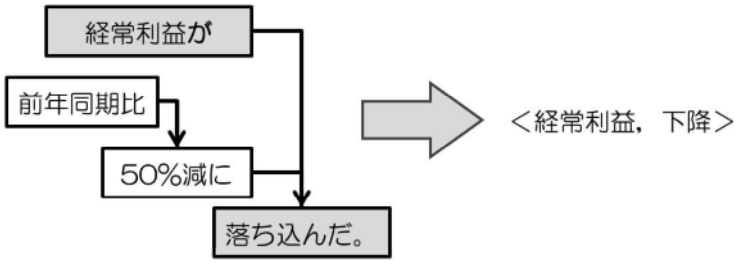

図 2: イベント表現の抽出例

VPを「上昇，下降，発生」に分類する方法については，本研究で作成した辞書を用いて行う．辞書の作成方法については 3.2.1 節で述べる．

3.2.1 辞書の作成方法

判定に用いる辞書は，WordNet を参考に人手で辞書を作成する．この辞書の語句数が多ければ多いほど，多くのイベント表現を抽出できることになる．そこで，上昇，下降の辞書については自動的に拡張を行い，最終的には拡張した辞書から人手で適切でない表現を除いた辞書を，VPの分類に用いる．

自動的な辞書の拡張は，人手で作成した辞書を種表現として，以下の 3 つの手順により行う．

手順 1 上昇，下降のいずれかの属性を持つ NP を種表現を用いて収集する

手順 2 手順 1 で収集された名詞句に係る VP を収集し，自己相互情報量 PMI を用いてフィルタリングを行う

手順 3 手順 2 で収集された表現について種表現と時間情報を用いて属性のタグ付けする

以下，それぞれの手順の詳細について述べる．

手順 1

ここでは，上昇，及び下降といった変化を持つ可能性のある NP を収集することを目的としている．そこで，人手で作成した辞書を種表現として，上昇，下降の種表現に係るすべての NP を，変化の対象となる NP として収集する．

手順 2

手順 1 で収集された NP が係る動詞，またはサ変接続の名詞から始まる文節 VP を収集する．そして，この文節から助動詞「ない」を除く助動詞，助詞，記号，数字を除去し，残った動詞についても原型に変換する．例えば,「供給されて,」という文節は,「供給するれる」と変換される．これは，上記の例や「激化したが,」という

ように，文節末に助詞や記号などが含まれていても，同じ語句として扱えるようにするために行う．そして,「ない」や「する」といった明らかに変化を表さないVPを，自己相互情報量PMIを用いて除外する．NPとVPのPMIの値が閾値βより小さい場合，そのVPを変化を表さない一般的な語句と見なして候補から取り除く（図3）．PMIは(1)式で計算され，辞書に加える条件は(2)式となる．

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (1)$$

$$PMI(w_i, w_j) > \beta \quad (2)$$

ただし，$p(w_i)$は語w_iの出現確率，$p(w_i, w_j)$は語w_iとw_jの同時出現確率を表す．

手順3

手順2で収集されたそれぞれのVPに対して,「上昇，下降，その他」のいずれかのタグを付与する．あるNPが種表現に係る文の表すイベントの発生時期の前後3日以内に，そのNPが手順2で収集されたVPに係っていた場合，そのVPに対してNPが係っていた種表現と同じ属性を付与する．そして，上記の方法でタグ付けされなかったVPは，その他とする．もし，このイベントの発生時期の前後3日以内に複数の属性の種表現に係っていた場合，多数決により収集したVPの属性を決定した．ここで同数であった場合は,「その他」のタグを付与した．

語Aが語Bに係っていた場合を「A → B」と表記すると，例えば「競争→増す」というイベントの発生時期と同時期に「競争→激化」というイベントが述べられている場合,「激化」は「上昇」の意味を持つとタグを付ける．

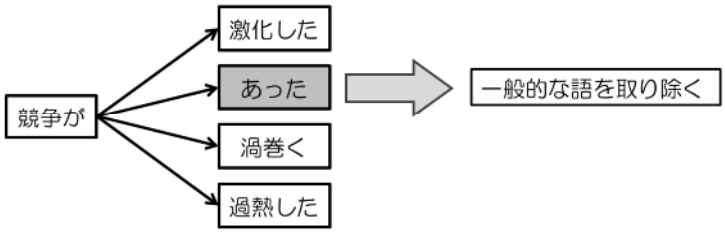

図 3: フィルタリングのイメージ

3.3 イベント表現のクラスタリング

抽出されたイベントの NP は，同じ意味であっても違う表記のものがある．そこで，同じ意味である NP については，表記が違っていても同じ NP として扱えるように，NP の同義語判定を行う．この同義語判定は，同じ意味の単語で構成されているかどうかに着目して行う．次に，同じ $\langle NP, VP \rangle$ で構成された，それぞれのイベント表現集合に対してクラスタリングを行う．これは，図 1 で挙げた例 ⟨ 事故, 発生 ⟩ のように，同じイベント表現が異なる話題を表すことがあるからである．

3.3.1 NP の同義語判定

抽出されたイベント表現の NP には,「資金ニーズ」と「資金需要」,「原油相場」と「原油価格」というように同じ意味であっても，違う表記のものがある．このような NP を同一の NP として扱えるようにするため，NP の同義語判定を行う．これは，WordNet を用いて意味的に同じ単語で構成されているかどうかに着目して行う．

具体的には，ある 2 つの名詞句，NP_1 と NP_2 があった場合，NP_1,NP_2 をそれぞれ単語の集合と考え，以下の手順で行う．

手順 1　NP_1 と NP_2 から，記号,「など」，接尾辞（～量，～率など）を削除する．

手順 2　NP_1 と NP_2 の両方に存在する単語と，2 つの集合間で WordNet を用いて同義語と判定される単語のペアを NP_1，NP_2 から削除する．

手順 3　NP_1 と NP_2 がともに空集合となった場合，NP_1 と NP_2 は意味的に同じ単語で構成されているといえるため，同義語と判定する．

同一と判断された NP は，すべて同一と判断された NP 内のいずれかの NP と置き換えて扱う．例えば，語 t_1, t_2, t_3 があり，同義語判定により $t_1 = t_2$，$t_2 = t_3$ と判断された場合，語 t_1, t_2, t_3 はすべて t_1 とみなして，以後の処理を行う．

同義語判定の処理の例として，2 つの名詞句,「原材料輸入価格」と「輸入原料価格など」に対して判定を行った場合の処理を図 4 に示す．

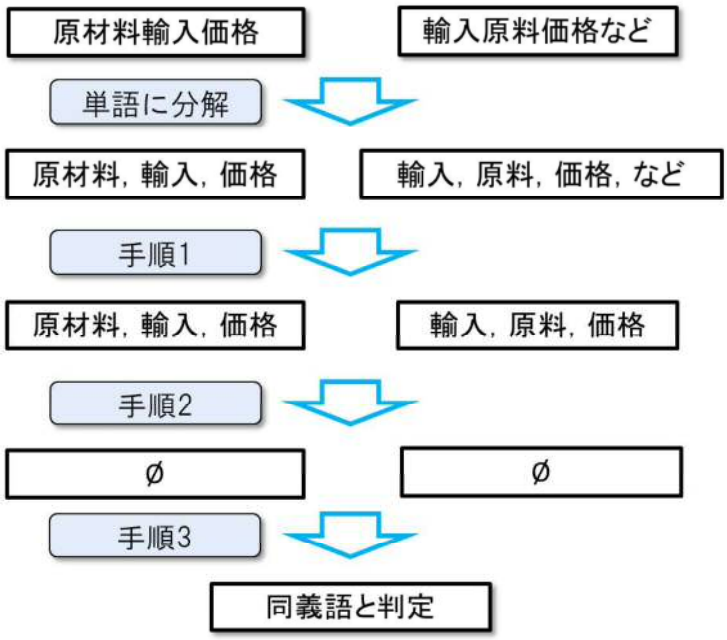

図 4: 同義語判定の実行例

3.3.2 $\langle NP, VP \rangle$ のクラスタリング

同じイベント表現であっても，属する話題が異なっていた場合，それらは別のものとして扱う方が適切である．例えば，同じ ⟨ 事故, 発生 ⟩ というイベント表現集合の中に交通事故の話題や，工場事故の話題などが混ざっていた場合，それらを同じ 1 つのイベントとして扱うのは不適切だと考えられる．そこで，それぞれのイベント表現 $\langle NP, VP \rangle$ 集合に対して，話題で分類するためのクラスタリングを行う（図 5）．ここで得られるそれぞれのクラスタが，因果関係知識における 1 つの事象となる．

ここではまず，出現したそれぞれのイベント表現のベクトルを生成する．イベント表現ベクトルは，イベント表現が含まれる文書に出現する語 t_i を要素として表現する．この語 t_i は名詞を対象とし，名詞が連続して出現していた場合は，それらを連結して 1 つの語として扱う．ベクトルの各要素は (3) 式によって決定され，イベン

ト表現は (4) 式のように表現される．

$$w_e(t_i) = tf_d(t_i) \times \log_2 \frac{|D|}{df_D(t_i)} \times \frac{1}{1 + \log(dist_e(t_i))} \quad (3)$$

$$\mathbf{e} = (w(t_1), w(t_2), \ldots, w(t_n)) \quad (4)$$

ただし，$tf_d(t_i)$ はイベント表現 e を含む文章 d 内での語 t_i の出現頻度，$|D|$ は総文書数，$df_D(t_i)$ は語 t_i の出現する文書数，$dist_e(t_i)$ はイベント表現 e を含む文と語 t_i を含む文の距離（2 文間に存在する文数+1）を表す．この重み付けは tf-idf 法をベースとして，イベント表現に近い単語はそのイベント表現が属する話題に関する語が記述されやすいという特性を考慮している．

上記の重み付け方法で生成したイベント表現ベクトルのままでは，比較するイベント表現ベクトルどうしが同じ単語を用いていなければ類似度が低くなってしまう．そこで，単語-イベント表現行列 M を作成し，その行列 M に対して次元圧縮を行う．本手法では，LSA（潜在的意味解析）により次元圧縮を行う．そのため，まず行列 M を特異値分解により

$$M = U\Sigma V^T \quad (5)$$

と分解する．そして，U の最初の k 個の左特異ベクトルのみから構成される U_k，k 個の大きな特異値のみから構成される Σ_k，最初の k 個の右特異ベクトルから構成される V_k を用いて，行列 M はランク k の行列 M_k に近似できる．

$$M_k = U_k \Sigma_k V_k^T \quad (6)$$

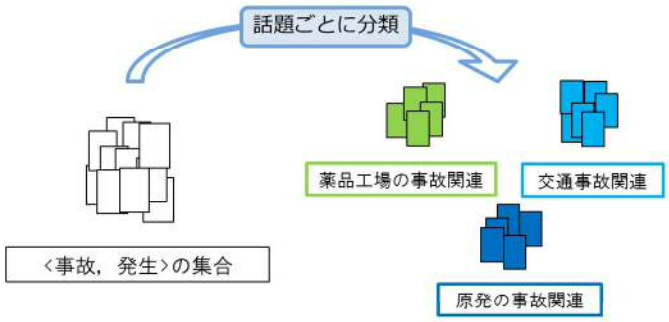

図 5: クラスタリングのイメージ

この行列 V_k の i 番目の要素をイベント表現 e_i の特徴ベクトル v_i^s として用いる．この特徴ベクトルを利用して，イベント表現間 v_i^s，v_j^s の類似度は，cos 類似度により計算する．

$$sim(e_i, e_j) = cos(v_i^s, v_j^s) = \frac{v_i^s \cdot v_j^s}{|v_i^s||v_j^s|} \tag{7}$$

そして，上記で述べた特徴ベクトルを用いてクラスタリングを行う．クラスタリングには，階層的クラスタリング手法である ward 法により行う．そのアルゴリズムを以下に示す．なお，クラスタリングの停止条件については，予備調査により決定した．

1. 各要素をそれぞれ要素数 1 のクラスタとする．
2. クラスタどうしのクラスタ間距離を求める．
3. クラスタ間距離が最も小さいクラスタを併合する．
4. 併合したクラスタ間の距離と，次に併合するクラスタ間の距離の差が 0.5 を超えた場合，終了する．超えていなければ，手順 2 に戻る．

3.4　時系列順への並び替えとバースト検出

2.2 節で生成された各クラスタごとに，そのクラスタに含まれる記事を時系列順に並び替える．そして，このイベント表現の時系列データに対してバースト検出を行う．バーストとは，ある時からある話題に関する記述が急激に増加するような現象を指す．バースト検出を行うことにより，イベントが話題になった時期がより鮮明になった時系列データを得ることができる（図 6）．

3.4.1　バースト モデル

本研究では，Kleinberg[14] のバースト検出アルゴリズムの 1 つである Enumerating バーストを用いる．これは，バースト解析において代表的なアルゴリズムである．Enumerating バーストのアルゴリズムは，離散時間で送られる集合に対して適用される．なお，本手法では 1 週間を単位とする．

解析期間において n 個のあるイベント表現 E の集合 $e_1, ..., e_n$ と，n 個の記事集合 $d_1, ..., d_n$ が離散時間で送られてくる状況を考える．解析期間における全てのイベン

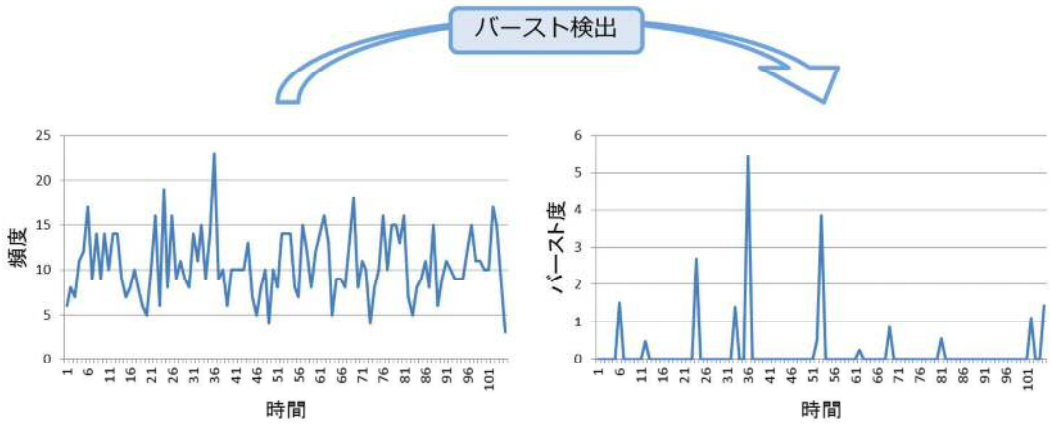

図 6: バースト検出のイメージ

ト表現 E の数 E_{all} は (8) 式，すべての記事の数 D は (9) 式と表すことができる．

$$E_{all} = \sum_{t=1}^{n} e_t \tag{8}$$

$$D = \sum_{t=1}^{n} d_t \tag{9}$$

次に，2 つの状態にそれぞれ期待値を割り当てる．非バースト状態 q_0 には，分析期間全体を見たときの期待値 $p_0 = E_{all}/D$ を割り当てる．バースト状態 q_1 には，p_0 にパラメータ s をかけた値である $p_1 = sp_0$ を割り当てる．ただし，$s > 1$ であり，$p_1 \leq 1$ となるような s でなくてはならない．記事集合中にイベント表現が二項分布 $B(d_t, p_i)$ にしたがって現れるという考えに基づき，状態 q_i にいることに対してコストを与える関数 $\sigma(i, e_i, d_t)$ は (10) 式のように定義される．

$$\sigma(i, e_t, d_t) = -In\left[\binom{d_t}{e_t} p_i^{e_t}(1-p_i)^{d_t - e_t}\right] \tag{10}$$

そして，期間 $t_k, ..., t_l$ におけるイベント e のバースト度 $b(t_k, t_l, w)$ は (11) 式の式で定義される．

$$b(t_k, t_l, e) = \sum_{t=t_k}^{t_l} (\sigma(0, e_t, d_t) - \sigma(1, e_t, d_t)) \tag{11}$$

本手法では 1 週間ごとにバースト度を算出するため，$t_k = t_l$ となる．

3.5 因果性の有無の判断

3.4 節で生成された，縦軸をバースト度としたイベントの時系列データ対を対象に，グレンジャー因果性検定を用いて因果性の有無を判断する．そして，検定で有意となったイベント対を因果関係知識として抽出する．

3.5.1 グレンジャー因果性検定

グレンジャー因果性とは経済学における時系列分析でよく用いられる概念であり，以下のように定義される [15].

グレンジャー因果性

現在と過去の x の値だけに基づいた将来の x の予測と，現在と過去の x と y の値に基づいた将来の x の予測を比較して，後者の MSE の方が小さくなる場合，y_t から x_t へのグレンジャー因果性 (Granger causality) が存在するといわれる．

グレンジャー因果性検定は，因果関係の有無を因果関係の有無を背後の明確な理論を用いずに，データだけから判断をしたいときに用いられる．この検定を行うために，因果性の有無を調べるイベントの時系列データ対を用いて，ベクトル自己回帰 (VAR) モデルの推定を行う必要がある．

VAR(p) モデルは，$\underline{\mathrm{y}}_t$ を定数と自身の p 期の過去の値に回帰したモデルである．本研究では 2 つのイベント間の因果性を調べるため，2 変量 VAR(p) モデルを具体的に表現すると (12) 式となる．

$$\begin{cases} y_{1t} = c_1 + \Phi_{11}^{(1)} y_{1,t-1} + \Phi_{12}^{(1)} y_{2,t-1} + \cdots \\ \qquad\qquad + \Phi_{11}^{(n)} y_{1,t-p} + \Phi_{12}^{(n)} y_{2,t-p} + \epsilon_{1t} \\ y_{2t} = c_2 + \Phi_{21}^{(1)} y_{1,t-1} + \Phi_{22}^{(1)} y_{2,t-1} + \cdots \\ \qquad\qquad + \Phi_{21}^{(n)} y_{1,t-p} + \Phi_{22}^{(n)} y_{2,t-p} + \epsilon_{2t} \end{cases} \tag{12}$$

この VAR モデルの推定は，各方程式を個別に最小二乗法 (OLS) によって行う．そして，VAR モデルを推定する際の次数 p は，AIC により選択する．この次数 p は，どの程度の過去まで考慮するかを表す．

グレンジャー因果性検定は，F 検定を用いてグレンジャー因果性を検定することで行う．グレンジャー因果性を検定するためには，$\Phi_{12}^1 = \Phi_{12}^2 = \ldots = 0$ を検定すれ

ばよい．そこで，まず

$$y_{1t} = c_1 + \Phi_{11}^{(1)} y_{1,t-1} + \Phi_{12}^{(1)} y_{2,t-1} + \cdots + \Phi_{11}^{(n)} y_{1,t-p} + \Phi_{12}^{(n)} y_{2,t-p} + \epsilon_{1t}$$

を OLS で推定し，その残差平方和を SSR_1 とする．次に，制約を課したモデル

$$y_{1t} = c_1 + \Phi_{11}^{(1)} y_{1,t-1} + \cdots + \Phi_{11}^{(n)} y_{1,t-p} + \epsilon_{1t}$$

を OLS で推定し，その残差平方和を SSR_0 とする．このとき，F 検定量は

$$F \equiv \frac{(SSR_0 - SSR_1)/2}{SSR_1/(T-5)} \tag{13}$$

で定義される．そして，$2F$ の値を $\chi^2(2)$ の 95%点と比較して，$2F$ の方が大きければ，y_{2t} から y_{1t} へのグレンジャー因果性が存在しないという帰無仮説を棄却し，y_{2t} は y_{1t} の将来を予測するのに有用であると結論する．

4 実験および評価

4.1 実験材料

用意した新聞記事は，1996∼1997 年に発行された毎日新聞，日経新聞，読売新聞の記事 890,041 件である．この記事集合に対して提案手法を適用し，3.2.1 節で述べた辞書の拡張と，因果関係知識の抽出を行った．なお，係り受け解析器には CaboCha[1]，VAR モデルの推定とグレンジャー因果性検定には R 言語のパッケージである $vars$[2] を用いた．

VP の分類に用いる辞書については，表 1 に示す語句を用意した．そして，表 1 に示した上昇，下降の表現を種表現として辞書の拡張を行った．なお，閾値 β は 5 とした．

因果関係知識を抽出する際には，VP の分類に，拡張された辞書を用いた．LSA によるイベント表現ベクトルの次元圧縮の際，次元数は 300 とした．そして，VAR モデルの推定では最大の次数を 8 とした．本手法では 1 週間を単位としているため，これは最大 2ヶ月のラグを考慮した推定となる．また，バースト検出におけるパラメータ s は $s = 2$ とし，グレンジャー因果性検定は有意水準は 5%で行った．対象としたイベントは，3.3 節で述べた方法でクラスタリングを行った結果，クラスタの要素数が 100 を超えたものにした．

表 1: 用意した辞書

属性	語句
上昇	増加，増進，増す，上昇，高まる，前進，上がる，アップ，上る，膨れ上がる，回復，再起，復活，復調
下降	下降，減少，縮小，下落，落ちる，落ち込む，ダウン，低減，低落，悪化，深刻化，低下
発生	起こる，出る，生じる，生起，発生，発する

[1] http://code.google.com/p/cabocha/
[2] http://cran.r-project.org/web/packages/vars/

4.2 結果

まず，辞書の拡張結果について述べる．自動的な拡張では，上昇の語句は 1175 個，下降の語句は 444 個収集された．その一部を，表 2 に示す．そして，自動的に収集された語句から，人手で適切でない表現を除いた結果，追加された上昇，下降の辞書の語数はそれぞれ 553 個，184 個となった．

次に，因果関係知識の抽出結果について述べる．要素数が 100 を超えたクラスタは 367 個であったため，解析対象のイベントは 367 種類となった．そして，本手法により抽出された因果関係知識は 7431 個であった．

4.3 評価

提案手法により獲得した因果関係知識の妥当性や新規性を，アンケートにより評価した．被験者は大学生 5 人である．

4.3.1 評価方法

獲得された因果関係知識 7431 件の中からランダムで 40 件抽出し，それらの知識の因果関係としての妥当性を評価してもらった．評価では，各知識に対して「因果関係である」,「因果関係でない」,「因果関係かどうかわからない」,「抽出されたイベントが理解できない」のどれかを選択してもらい，過半数の人が選択した項目を採用して集計した．

表 2: 自動的に収集された語句

属性	表現
上昇	上向きにある，活発化，浮く，こだまする，重要になる， クローズアップされる，熱中する，入れ替わる，外れる 得られやすい，はずむ，膨張する，食い違う，起きていない
下降	発揮できない，急落する，変色する，空回りする，ちらつく， 下方修正する，持ち直す，更新する，ひどい，急反落する， 重要でなくなる，赤字，氾濫する，正常化する，売買する

因果関係であるかどうかの判断は，乾ら [11] の方法と同様に，言語プレート（表3）を用いた方法で行なってもらった．言語プレートを用いることで，被験者間に共通の，言語テンプレートという言語的な判断の拠り所が与えられる．なお，言語プレートにおける $\langle\langle adv \rangle\rangle$ 部分について，乾らは「しばしば」や「大抵」などを代入して因果関係の強さを考慮するために用いていたが，本研究では考慮しないため無視した．

因果関係の有無の判断は，次の手順により行なってもらった．

1. それぞれのテンプレートに対して，因果関係の有無の判断を行う対象の原因部分を e_1，結果部分を e_2 に代入する．

2. 完成したテンプレート文の文意が意味的に適格であるかどうかを考える．

3. もし，あるプレート文が意味的に適格であると判断されれば，スロットに代入した2つのイベント間には因果関係があると判断する．

4. もし，18通りのすべてのテンプレート文がいずれにおいても適格と判断されない場合，2つのイベント間には因果関係がないと判断する．

表 3: 乾らが用いた言語プレート

id	the linguistic templates
1	『e_1』(という) ことが起こるその結果として、≪adv≫『e_2』(という) ことが起こる。
2	『e_1』(という) 状態になれば、それに伴い、≪adv≫『e_2』(という) 状態になる。
3	『e_1』(という) 状態になれば、それに伴い、≪adv≫『e_2』(という) 状況になる。
4	『e_1』(という) 状態であると、≪adv≫ 『e_2』(という) 状態である。
5	『e_1』(という) 状態であると、≪adv≫ 『e_2』(という) 状況である。
6	『e_1』(という) ことをする結果、『e_2』(という) ことが ≪adv≫ 起こる。
7	『e_1』(という) ことをすると、 ≪adv≫ 『e_2』(という) 状態になる。
8	『e_1』(という) ことをすると、 ≪adv≫ 『e_2』(という) 状況になる。
9	『e_1』(という) ことをすると、≪adv≫ 『e_2』(という) 状態を保つ。
10	『e_2』(という) ことをするのは、≪adv≫ 『e_1』(という) 状態の時である。
11	『e_2』(という) ことをするのは、≪adv≫ 『e_1』(という) 状況の時である。
12	『e_1』(という) 状態になる場合、≪adv≫ 『e_2』(という) ことをする。
13	『e_1』(という) 状況になる場合、≪adv≫ 『e_2』(という) ことをする。
14	『e_1』(という) 状態では、≪adv≫ 『e_2』(という) ことをする。
15	『e_1』(という) 状況では、≪adv≫ 『e_2』(という) ことをする。
16	『e_1』(という) ことが起こらなければ、≪adv≫ 、『e_2』(という) ことができない。
17	X が『e_2』(という) ことを実現する手段として、≪adv≫ X が『e_1』(という) ことを行なう。
18	X が『e_1』(という) ことをすることによって、≪adv≫ X が『e_2』(という) ことができる。

表 4: イベントの表現方法

ラベル	NP	属性
レベル, 選手, チーム, 大会, 試合, 優勝, 順位, プレー, 練習, 五輪	レベル	上昇
採算, %減, 悪化, 経常利益, 売上高, 円, 減少, 営業利益, %, 中間期	採算	下降
症状, O 1 5 7 , 下痢, 検出, 血便, 菌, 女児, 感染, 腹痛, 食中毒	症状	発生

イベント表現の話題については，クラスタ内の単語（名詞）の重要度を計算し，その上位 10 件を提示した．このイベントの話題を表す 10 個の単語を，本研究ではラベルと呼ぶ．よってイベントは表 4 のように表現される．

単語の重要度の計算方法

それぞれのクラスタは，(4) 式で表されるイベント表現ベクトル集合とみなすことができる．そして，あるクラスタ $C = (\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}, \ldots, \mathbf{e_n})$ があった場合，このクラスタ C 内にある単語 t_i の重要度 $w_C(t_i)$ は，それぞれのイベント表現ベクトルの値を (14) 式により正規化した後，(15) 式により計算される．

$$e_j^N(t_i) = \frac{e_j(t_i)}{\sum_{i=1}^{m} e_j(t_i)} \tag{14}$$

$$w_C(t_i) = \sum_{j=1}^{n} e_j^N(t_i) \tag{15}$$

ただし，m はイベント表現ベクトル e_j 内における単語の種類の数，$e_j(t_i)$ はイベント表現ベクトル e_j における単語 t_i の重みを示す．

4.3.2　評価結果

アンケートの結果は，以下の通りになった．

- 因果関係がある　　：20 件
- 因果関係がない　　：12 件
- 因果関係かどうかわからない：1 件
- イベントが理解できない　　：1 件

残りの 6 件は，過半数の被験者が選択した項目がなかった．

因果関係であると判断された知識と因果関係でないと判断された知識，そして因果関係かどうかわからないと判断された知識の一部を表 5 に示す．なお，表 5 に示す因果関係知識は，ラベルをみて人手により整形したものである．

4.4　考察

4.4.1　辞書の拡張方法について

まず，辞書の拡張に伴う辞書の語数の変化をまとめたものを，表 6 に示す．なお，「採用された割合」は，辞書に追加した数を自動的に収集した数で割った商である．

表 5: アンケート結果の一部

アンケート結果	内容
因果関係あり	消費税率が上がると，政府への不満が増える． 収益が減ると，経営難になる． 新生党内の確執が表面化すると，自民党の発言力が増す．
因果関係なし	法案が有力になると，がんの危険性が上がる． 証券会社で支障が出ると，スポーツ界において成果が上がる． パソコンの機能が良くなると，地震が起こる可能性が上がる．
わからない	車業界の競争が激化すると，政府内で構想が練られる．

表 6: 辞書の語数の変化

属性	種表現の数	自動的に収集した数	辞書に追加した数	採用された割合 (%)
上昇	14	1175	553	47
下降	12	444	184	41
合計	26	1619	739	46

この結果より，人手で用意する辞書の語句数は少数であっても，本手法を用いることで大幅に拡張できることがわかる．また，上昇や下降の属性を表す語句を，情報がゼロの状態から人手で数百個用意することは非常にコストのかかる作業であるが，すでにその候補が提示された状態から選択する作業は上記の作業に比べて容易である．実際に，上昇の属性における「エスカレートする」や「台頭する」，下降の属性における「続落」や「減退する」といった，人では容易に想像できない語句を自動的に収集することができていた．よって，VP の分類に用いる辞書の拡張方法として，本手法は有効であるといえる．

今後の課題として，主に 2 点挙げられる．1 つ目は，自動的な収集の精度の向上である．収集の精度が高まれば，より低いコストで辞書の拡張を行うことができるようになる．しかし，本手法の自動的な収集の精度は 46%であり，自動的に収集された語句をそのまま VP の分類に用いるには十分な精度とはいえない．そこで，自動的に収集された適切でない語句を分析した．すると,「災いする」や「餓死する」など，文書で使われにくい語句が多くあった．これらの特徴として，これらの語句に係る語も，事件名や人名といった固有名詞が多く，文書で使われにくい語が使われやすいという点が挙げられる．よって，この NP と VP は PMI の値が高くなりやすく，3.2.1 節における手順 2 のフィルタリングで除かれなかったと考えられる．この問題を解決するための 1 つの方法として，3.2.1 節における手順 1 と手順 2 の間で，変化を持つ可能性のある NP に対して，適切でない VP を収集しやすい NP かどうかの判断を行うことが考えられる．この判断を行うことで，精度が向上すると思われる．

2つ目は，2文節以上にわたって属性が表現される語句への対応である．本手法では，1文節単位で収集を行なっているが，実際には「拍車がかかる」や「～の状態が続く」,「～に変わる」など，2文節以上にわたって属性を表している語句が存在していた．したがって，このような表現に対応した辞書の構築が必要であると考えられる．

4.4.2 獲得した因果関係知識について

評価結果より，提案手法を用いて因果関係があると判断される知識を抽出できていたことがわかる．これより，因果関係である事象対は文書内で共起しているという前提を用いなくとも，イベントの発生した時期を捉えることで因果関係知識を獲得できるといえる．

因果関係でないと判断された抽出知識には「法案が有力になると，がんの危険性が上がる」や「証券会社で支障が出ると，スポーツ界において成果が上がる」というような，原因と結果の話題間に整合性がないものが多かった．例えば，結果のイベントが「がんの危険性が上がる」であった場合，前提として原因の話題は政治ではなく環境や生活でないと話題間に整合性がなく，因果関係とはならない．したがって，妥当な因果関係知識を抽出するためには，イベント間の関係を発生した時期だけでなく，話題という観点から考慮する必要があると考えられる．

抽出された知識には，獲得した知識内に因果関係かどうかわからないと判断される知識が存在していた．因果関係であるかどうかは，言語プレート用いて判断してもらっているため，完全な主観ではないが，被験者の知識に左右される部分がある．よって，知識が因果関係かどうかわからないと判断されたのは，偶然被験者が知らなかったためとも考えられる．しかし，この結果は本手法を用いることで新たな知識発見となる因果関係知識を抽出できることを示唆しているといえる．

過半数の被験者が選択した項目がなかった知識が6件あった．この結果となった原因の1つに，ラベルがあると考えられる．ラベルは10個の単語で構成されているが，そのすべてが統一の話題に関する単語でないラベルがいくつか存在した．そのラベルの例を以下に示す．

- 意見, 委員, 大勢, 公共事業, 中医協, 導入, 住民, 国, アセス, 結論
- 私, 気, 犬ベル, 人ごと, 急, とよのさん, そば, 施政方針演説, 気配, 学校

上記のようなラベルでは，イベントの話題の判断が一意に定まらず，被験者によってイベントの話題の捉え方に違いが発生した可能性がある．そして，この話題の捉え方の違いから，因果関係の有無の判断が別れる因果関係知識が発生したと考えられる．したがって，今後の課題としてイベントの話題を表すラベルの生成方法の開発が挙げられる．

ここで，アンケートで因果関係があると判断された知識を，実験材料から手がかり表現を用いた手法で抽出を試みた．手がかり表現を用いた手法は，多くの既存研究で用いられている手法である．用いた手がかり表現は，乾ら [10] と青野ら [16] の調査結果を参考に,「に伴う」,「に伴い」,「を理由に」,「が理由で」,「ため」の 4 つを用いた．そして，あるイベント対「$\langle NP_1, VP_1 \rangle$が原因で$\langle NP_2, VP_2 \rangle$となる」が妥当と判断された知識内にあったとき，手がかり表現より前に NP_1 と VP_1 が存在し，かつ後ろに NP_2 と VP_2 がある文が存在する文を抽出した．

（NP_1, VP_1）＜手がかり表現＞（NP_2, VP_2）。

ただし，VPが「発生」の場合,「発生した」などと属性が明示的に述べられないことが多いため，その場合VPは存在していなくても抽出した．また,「これに伴い，～」というように，手がかり表現が文頭に存在した場合は，その手がかり表現を含む文の前の文から NP_1, VP_1 を探し，存在していれば抽出した．

その結果を，以下に示す．抽出された知識のイベントは,「ラベル/NP/属性」という形式で表している．

- 本手法で抽出された知識
  - 原因
    - パソコン, 情報, 障害者, 電源, インターネット, 瀬さん, パソコン通信, 倉田氏, 記憶, 主要部品/情報/上昇
  - 結果
    - 件, 可能性, 帝国データ, 着服, 倒産, 倒産件数, 被害者, 商工リサーチ, 認証取得, 破産件数/可能性/上昇
- 新聞記事から抽出された文章
  - マルチメディア社会では、インターネットに代表されるような電子メディア情報がますます多様化する。それに伴い情報の流通過程で社会的摩擦が増える可能性がある。

上記した本手法により抽出された知識と新聞記事から抽出された知識を，アンケートに回答してもらった被験者に見比べてもらい，同じ知識かどうかを判断してもらった．その結果，過半数の被験者が同じであると回答した．よって，20 件中 1 件の知識は手がかり表現を用いた手法で抽出することができたといえる．言い換えると，20 件中 19 件の知識は，手がかり表現を用いた手法で抽出することができなかった，ということである．この結果から，既存手法では獲得しづらかった知識も本手法は獲得できているといえる．

4.4.3 バースト検出の有効性について

VAR モデルで推定される次数がどのように変化するかを調べることによって，バースト検出が本手法においてどのような影響をもたらしているかについて考察する．獲得されたすべての知識について，VAR モデルの次数は図 7 のようになった．比較として，バースト検出を行わずに獲得された知識の次数分布も記載した．

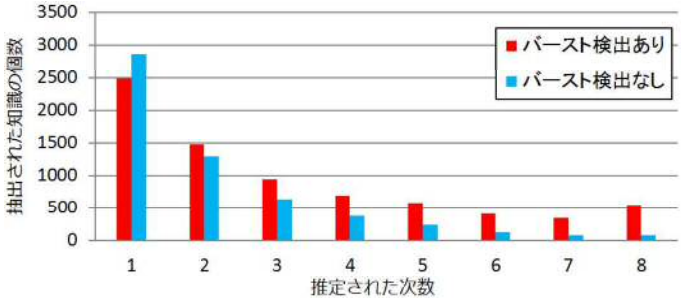

図 7: バースト検出の有無と推定された次数の関係

図 7 をみると，バースト検出を行うと高い次数の知識が抽出されやすいとわかる．つまり，バースト検出を行うことで，影響が出るまでのラグが大きいイベント対も因果関係知識として抽出されるということである．

これより，イベントの時系列データに対してバースト検出を行うことにより，人では気づきにくい因果関係知識を獲得できるといえる．

5　おわりに

本研究では，これまで多くの因果関係抽出の研究で用いられてきた，因果関係にある事象対は文書内で共起しやすいという前提を用いずに，イベントの時系列分析を行うことで因果関係知識を獲得する手法を提案した．

評価実験により，提案手法を用いることで妥当な因果関係とともに，既存手法では獲得しづらかった知識も獲得できることを確認できた．また，イベントの時系列データに対してバースト検出を行うことで，人では気づきにくい因果関係知識を獲得できることがわかった．

今後の課題として，イベントの話題を表すラベルの生成方法の改良や，話題という観点からイベント間の関係を考慮する方法の開発などが挙げられる．

参考文献

[1] Roxana Girju: Automatic detection of causal relations for question answering，*In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pp.7683 (2003).

[2] 佐藤　岳文，堀田　昌英: Webマイニングを用いた因果ネットワークの自動構築手法の開発，社会技術研究論文集，Vol.4，pp.66-74 (2006).

[3] 坂地泰紀，竹内康介，増山　繁，関根　聡：構文パターンを用いた因果関係の抽出，言語処理学会第 14 回年次大会論文集，pp.1144-1147 (2008).

[4] Khoo, C.S.G., Chan, S. and Niu, Y.: Extracting causal knowledge from a medical database using graphical patterns, *Proc. of the 38th ACL* , pp.336-343 (2000).

[5] Kentaro Torisawa: An unsupervised learning method for commonsensical inference rules on events，*In Proc. of the Second CoLogNet-ElsNET Symposium*，pp.146-153 (2003).

[6] Du-Seong Chang and Key-Sun Choi: Causal relation extraction using cue phrase and lexical pair probabilities，In *Proc. of the 1st IJCNLP*，pp.61-70 (2004).

[7] 山田　一郎，小早川　健，三浦　菊佳，住吉　英樹，八木　伸行，崔　杞鮮: クローズドキャプションを対象とした因果関係知識抽出の検討，In *FIT*，(2005).

[8] Mehwish Riaz and Roxana Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Proceedings of the 4th IEEE International Conference on Semantic Computin*，pp.361-368，(2010).

[9] Quang Xuan Do，Yee Seng Chan，and Dan Roth，Minimally supervised event causality identification: In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*，pp.294-303，(2011).

[10] 乾　考司，乾　健太郎，松本　裕治： 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得，情報処理学会論文誌，Vol.45 Mo.3　（2004）．

[11] 乾孝司, 奥村学: 文章内に現れる因果関係の出現特性調査，計量国語学, Vol.25, pp.123-144 (2005).

[12] Yizhou Sun，Kunqing Xie，Ning Liu，Shuicheng Yan，Benyu Zhang，and Zheng Chen: Causal relation of queries from temporal logs，In *Proceedings of the 16th International Conference on World Wide Web*，pp.1141-1142，(2007).

[13] 橋本泰一，村上浩司，乾孝司，内海和夫，石川正道: 社会課題発見のための文書クラスタリングとクラスタ評価指標，人工知能学会論文誌 Vol24, No.4 P333-338 (2009).

[14] Jon Kleinberg: Bursty and hierarchical structure in streams, *Data Mining and Knowledge Discovery*, Vol.7, No.4, pp.373-397 (2003).

[15] 沖本 竜義: 経済・ファイナンスデータの計量時系列分析，朝倉書店 (2010)．

[16] 青野 壮志，太田 学: 要因検索による因果関係ネットワークの構築と因果知識の獲得，*DEIM Forum 2010 B9-1* (2010).

謝辞

本研究を行うにあたり，丁寧な指導を頂きました内海彰教授に深く感謝致します．また，評価実験に実施にあたり，貴重な時間を割いて評価アンケートに協力して下さった皆様，日頃より多くの知識，示唆を頂いた内海研究室の皆様に感謝致します．

A　予備調査

A.1　イベント表現の抽出

イベント表現を抽出するにあたって，どのようにイベント表現を抽出すれば良いかについて調査した．

まず,「AがBした」,「AはBした」という情報を抽出したいため，動詞（サ変接続の名詞）と，動詞に直接係る格助詞「が，は」を含む文節の係り受け関係を1つのイベント表現し，抽出した．その結果を表7に示す．なお，これは新聞記事1年分（441193記事）から抽出した結果である．表7みると,「ある，いる」という一般的な語に係ることが多かった．そのため，これらの単語をストップワードとし，これらに係るペアを抽出の対象から外した．そして，同様に抽出を行った．その結果を表8に示す．

表8をみると,「分かる」や「大切」といったイベントを表す動詞としては適切でない表現が多く抽出されることがわかる．ここで，因果関係の事象となるイベントとはどのようなものかを考えた場合，それは何かが変化したイベント，または何かが起こったイベントであると思われる．

したがって，動詞に直接係る格助詞「が，は」を含む文節の係り受け関係で，かつ動詞にあたる部分が「上昇，下降，発生」に関する語句に着目するべきだという結論に至った．

表 7: ペアの抽出例

係り元	係り先	頻度
ことが	できる	12088
必要が	ある	10720
ことが	ある	8292
ことが	分かる	6465
可能性が	ある	5981
ことが	わかる	3821
ことが	明らか	3374
恐れが	ある	3127
気が	する	3061
ことが	多い	2976
可能性が	高い	2758
問題が	ある	2489
時間が	かかる	2427
ことが	必要	2261
ものが	ある	2176
ことが	ない	2056
ことが	なる	2051
人が	いる	2051
人が	多い	1949
方が	いい	1910
公算が	大きい	1730
見方が	多い	1716
ことが	判明	1620
ケースが	多い	1560
影響が	出る	1471
声が	出る	1403
買いが	入る	1377
ことが	決まる	1352
ことが	重要	1319

表 8: ストップワードを用いた抽出例

係り元	係り先	頻度
ことが	できる	12088
ことが	分かる	6465
ことが	わかる	3821
ことが	明らか	3374
気が	する	3061
ことが	多い	2976
可能性が	高い	2758
時間が	かかる	2427
ことが	必要	2261
ことが	ない	2056
ことが	なる	2051
人が	多い	1949
方が	いい	1910
公算が	大きい	1730
見方が	多い	1716
ことが	判明	1620
ケースが	多い	1560
影響が	出る	1471
声が	出る	1403
買いが	入る	1377
ことが	決まる	1352
ことが	重要	1319
ことが	大切	1317
声が	上がる	1271
声が	多い	1265
可能性が	強い	1227
ことが、	分かる	1128
動きが	出る	1076
動きが	広がる	1041

A.2　辞書の拡張方法

VPの分類に用いる辞書を自動的に拡張するにあたって，どのように上昇や下降を表すような語句を収集すれば良いかについて調査した．

まず,「株価」や「競争」など，上昇や下降といった変化を持つ可能性のあるNPが係る動詞，またはサ変接続の名詞を収集し，その頻度の高い語句を調べた．変化を持つ可能性があるNPには，上昇，下降，発生の種表現に係る名詞句を用いた．その結果は，以下のようになった．なお,「上昇」に付随する単語集合は，上昇の種表現を用いて得られたNPを用いて収集された語句である.「下降」と「発生」についても，同様である．

- 上昇
  - 多い, なる, ない, 強い, 高い, 増える, 必要, 大きい, 続く, する, できる, かかる, 入る, 目立つ, 進む, 少ない, 強まる, いい, 広がる, 相次ぐ, 増えるいる, 続くいる, なるいる, 集まる, 強まるいる, 起きる, 始まる, 減る, 占める, 広がるいる
- 下降
  - 多い, なる, ない, 高い, 強い, 増える, 大きい, かかる, 続く, する, 少ない, 入る, 必要, いい, できる, 目立つ, 増えるいる, 減る, 続くいる, 広がる, なるいる, 集まる, 伸びる, 強まる, つく, 低い, 占める, 進む, 悪い, 変わる
- 発生
  - 多い, なる, ない, 強い, する, 必要, 続く, 大きい, 高い, できる, 増える, 目立つ, 入る, かかる, 起きる, 進む, 広がる, 続くいる, 少ない, 相次ぐ, 強まる, いい, 集まる, なるいる, つく, 増えるいる, 強まるいる, 始まる, 残る, 広がるいる

上記の結果をみると，すべての属性において同じような語句が収集されていることがわかる．これは，語句の収集のために利用しているNPが，3つの属性間で重なっているためであると考えられた．

そこで，頻度のみを利用するのではなく，収集した語句が出現している時期も利用することで，辞書に追加する語句の収集を試みた．以下に示す結果は，ある NP が種表現に係る文の表すイベントの発生時期と，同じ日に，その NP が係る動詞，またはサ変接続の名詞を収集した．示した語句は，その中で頻度の高い語句である．

- 上昇
  - 相次ぐ, 贈るれる, 根強い, つく, 決まる, 相次ぐいる, 先行する, 死亡, 強まる, 膨らむ, かかる, 違う, 出すれる, 求める, 負う, 固まる, 進むいる, 大勢, 好き, 逮捕するれる, 決める, 見込むれる, 表面化する, 続出, 売るれる, 鳴る, 優れるいる, ため, 手渡すれる, 可能なる, 厚い
- 下降
  - いい, 大きい, よい, 良い, 拡大する, いいの, 有利, 安い, するいる, 開く, 広い, 下がる, 知る, 大きいの, 望ましい, 早い, 回る, いいん, 楽, 上, 流れる, 判断する, 急落する, 加わる, もの, 長い, 楽しめる, 襲うれる, 動く, 行方不明なるいる, 低迷するいる, 低迷する, 死傷する
- 発生
  - 多い, ない, なる, 占める, 増える, 高い, 目立つ, 必要, 続く, 強い, できる, 少ない, 得るられる, 続くいる, 達す, 寄せるられる, する, 多いの, 広がる, 超える, 増えるいる, 減る, 集まる, なるいる, 起きる, 残る, なくなる, 始まる, 上回る, 入る, 死亡する, 終わる, 進む

上記の結果をみると，上昇と下降については，頻度のみで収集した結果と比べて良くなっていることがわかる．しかし,「違う」や「いい」などの，変化を表さない一般語が含まれていることがわかる．そこで，NP と VP の PMI を計算し，一般語を取り除こうと考え，本手法に至った．

発生については大きな変化がみられず，また発生の属性を表す語句もみられなかったため，自動的な拡張は困難であると判断し，行わないことにした．

A.3　クラスタリングの停止条件

イベント表現$\langle NP, VP\rangle$のクラスタリングを行う際の，停止条件について調査した．

このクラスタリングの目的は，イベント表現を話題ごとに分類することである．そこで，基準値（併合されたクラスタ間の距離）を分析した．異なる話題のクラスタを併合するとき，この基準値が急激に大きくなると考えられるためである．図8~13に，300~350個のイベント表現集合のクラスタリングにおける基準値の推移を示す．また，図14~19に700~750個のイベント表現集合のクラスタリングにおける基準値の推移を示す．

図8~19をみると，急激に基準値が上昇する値は，イベント表現の数に依存して変化することがわかる．そこで，直前の基準値との差を調べた．図20~25に，300~350個のイベント表現集合のクラスタリングにおける直前の基準値との差の推移を示す．また，図26~31に700~750個のイベント表現集合のクラスタリングにおける直前の基準値との差の推移を示す．

図20~31をみると，イベント表現の数に関わらず，基準値の差が0.5を超えたときから差が大きくなっていることがわかる．

そこで本手法では，クラスタリングの停止条件として，次の基準値との差が0.5を超えた時点とした．

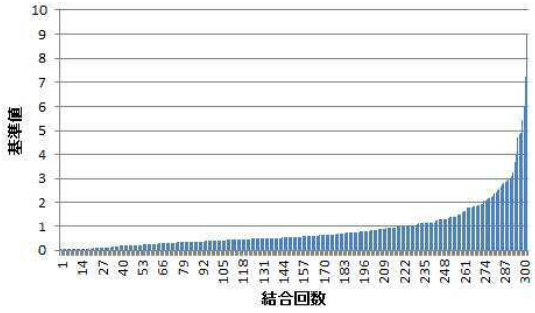

図 8: 基準値の推移 (a)

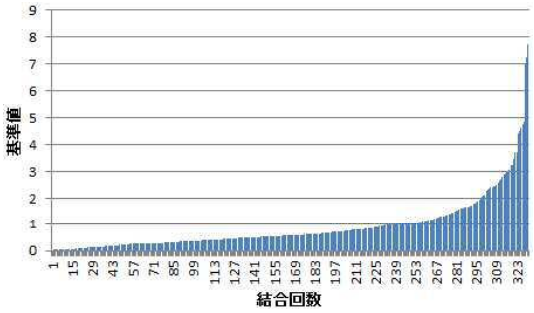

図 9: 基準値の推移 (b)

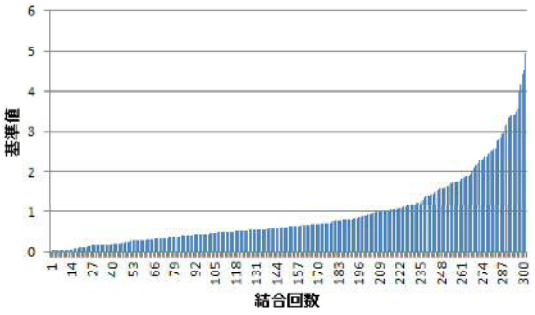

図 10: 基準値の推移 (c)

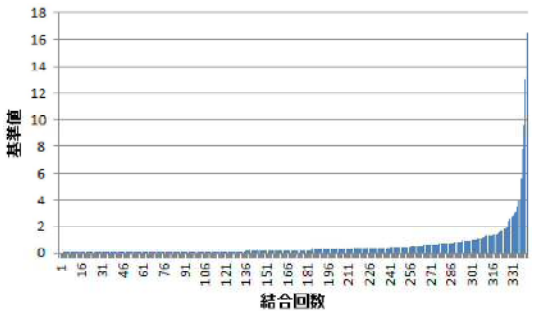

図 11: 基準値の推移 (d)

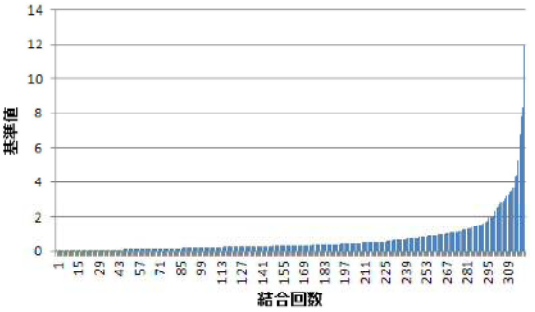

図 12: 基準値の推移 (e)

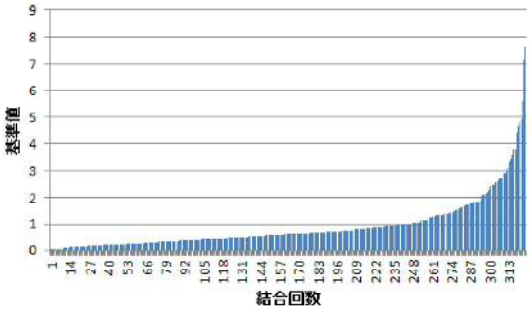

図 13: 基準値の推移 (f)

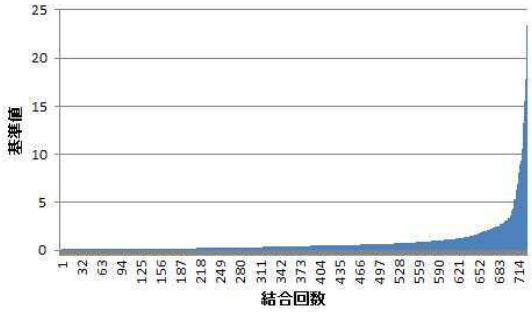

図 14: 基準値の推移 (g)

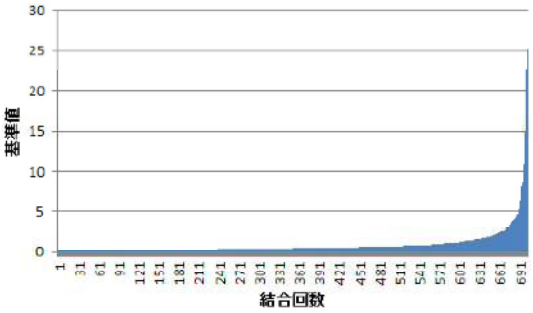

図 15: 基準値の推移 (h)

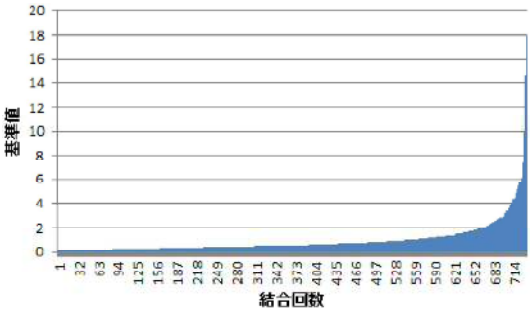

図 16: 基準値の推移 (i)

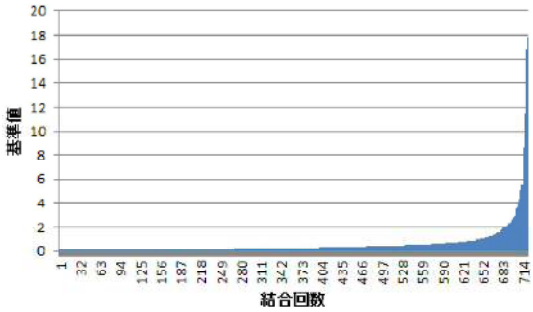

図 17: 基準値の推移 (j)

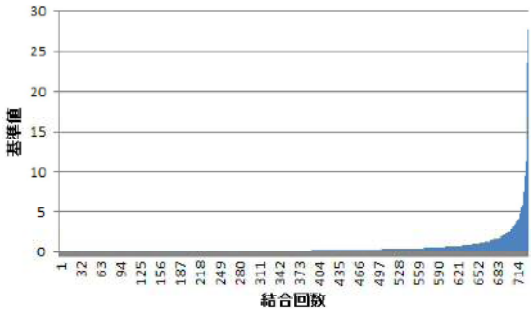

図 18: 基準値の推移 (k)

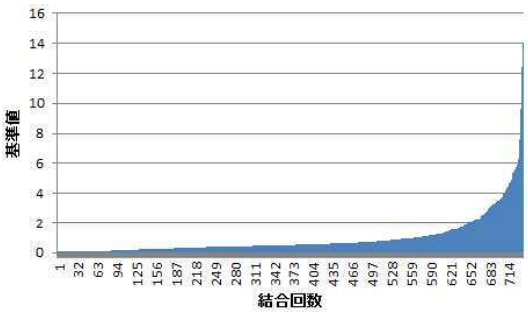

図 19: 基準値の推移 (l)

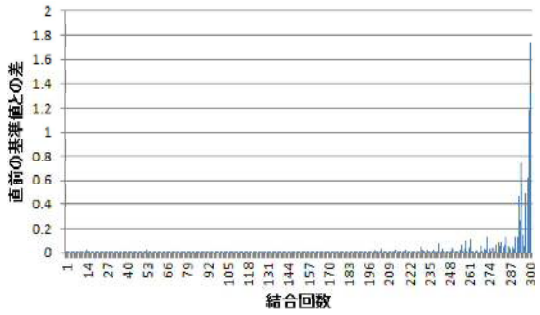

図 20: 基準値の差の推移 (a)

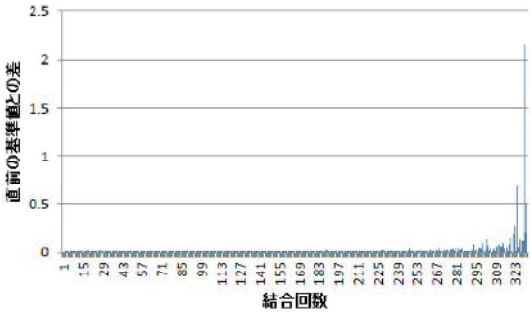

図 21: 基準値の差の推移 (b)

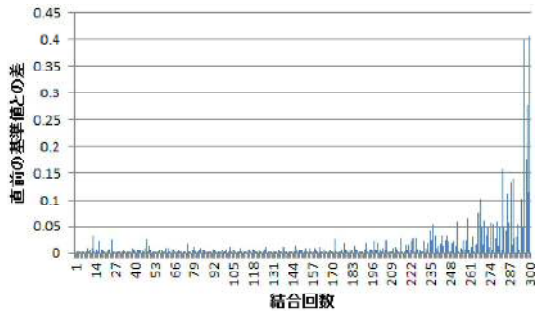

図 22: 基準値の差の推移 (c)

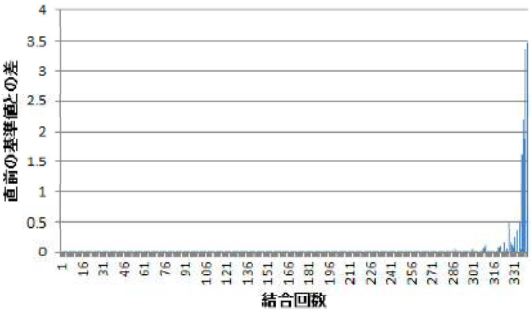

図 23: 基準値の差の推移 (d)

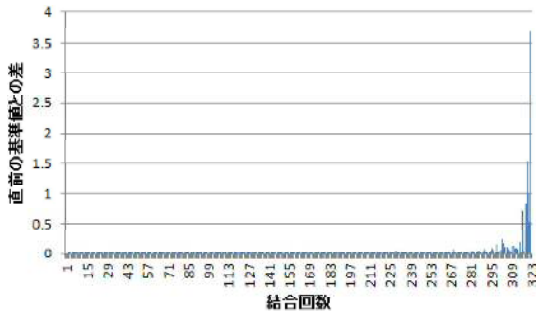

図 24: 基準値の差の推移 (e)

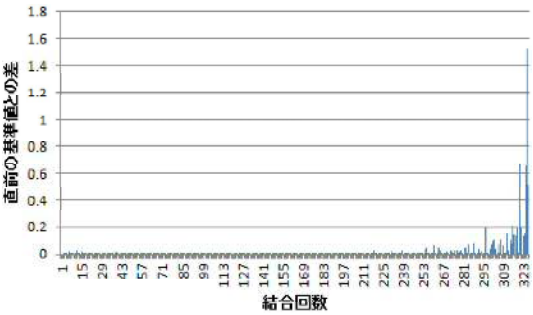

図 25: 基準値の差の推移 (f)

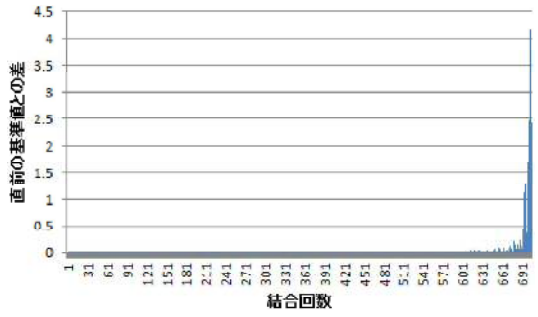

図 26: 基準値の差の推移 (g)

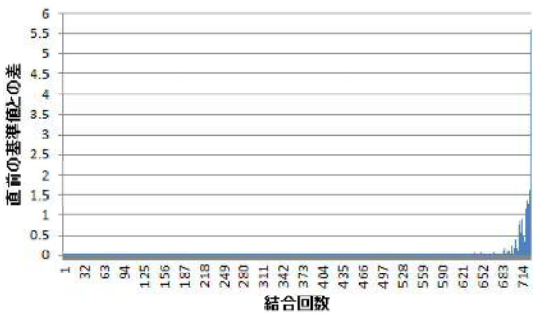

図 27: 基準値の差の推移 (h)

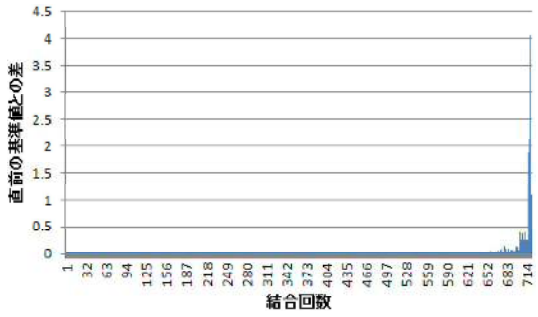

図 28: 基準値の差の推移 (i)

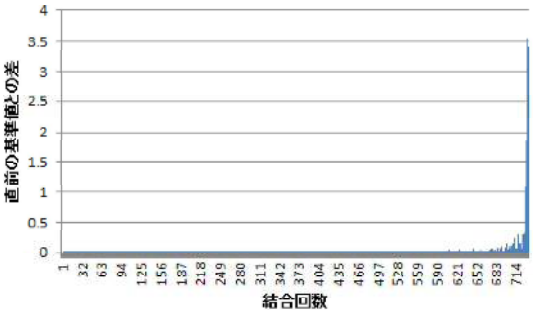

図 29: 基準値の差の推移 (j)

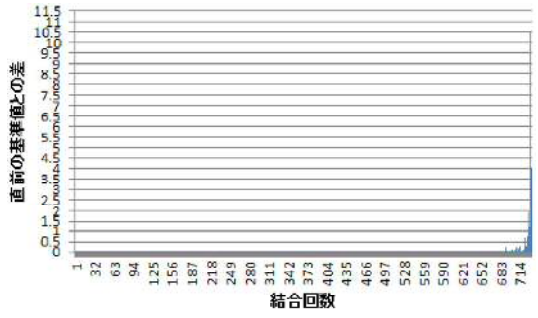

図 30: 基準値の差の推移 (k)

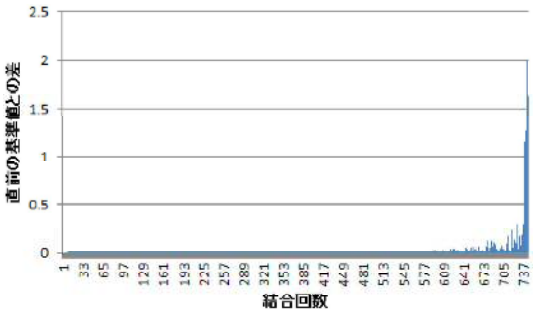

図 31: 基準値の差の推移 (l)

B　アンケート結果

評価において行ったアンケートの結果を示す．因果関係があると判断された知識を表 9 と表 10，因果関係がないと判断された知識を表 11，過半数の被験者が選択した項目がなかった知識を表 12 に示す．また，因果関係かどうかわからないと判断された知識と，イベントが理解できないと判断された知識を以下に示す．この 2 つの知識については,「ラベル/NP/属性」という形式で表している．

- 因果関係かどうかわからないと判断された知識
  - 原因
    - 競争, 激化, 台, 生産, 携帯電話, 競合, マツダ, 開発, トヨタ, フォード / 競争/上昇
  - 結果
    - 構想, 浮上, 社民党, 政府, 自民党内, 空中権, 与党内, 党, 与党, 酒造会社/構想/上昇
- イベントが理解できないと判断された知識
  - 原因
    - 方, 勇み足, 高畑, ん, 笑い, ジャンプスーツ, 庄田, ライン, マナー, プレッシャー/方/発生
  - 結果
    - ドイツ, スト, 反発, ルフトハンザ, フランス, 通貨統合, 労組, 労働時間, 独仏, シラク政権/反発/上昇

表 9: 因果関係があると判断された知識 (a)

原因			結果		
ラベル	NP	属性	ラベル	NP	属性
消費税率, %, 税金, アップ, 四月, 特別減税, 国民, 消費税, ４月, 来年	消費税率	上昇	首相, 不満, 省庁, 行革会議, 委員, 議論, 総務会, 中間報告, 噴出, クラウス首相	不満	上昇
トン, 可能性, 在庫, コメ在庫, 予測修正率, 中国, 豊作, 年産米, 生産, 今後	可能性	上昇	見方, 値上げ, トン, 市中価格, 在庫, 有力, 上昇, 輸入量, 減産, 需要家	見方	上昇
低下, 機能, 軸封部, かん, 販売店, ブレンダー食ミニ, 洗浄液, 脳疾患患者, 北陸電力, 三和化学研究所	機能	下降	声, 社長, 株主, 責任, 会長, 若狭氏, 退任, 辞任, 総会, 社内	声	上昇
人気, 水草,—, 円, 河童, シーキングザパール, ラフティング, 魚, 水, キロ	人気	上昇	生産, 工場, 産地, 生産量, 生産性, 回復, 岐阜工場, 増加, 国産合板, 下期	生産	上昇
パソコン, 情報, 障害者, 電源, インターネット, 瀬さん, パソコン通信, 倉田氏, 記憶, 主要部品	情報	上昇	件, 可能性, 帝国データ, 着服, 倒産, 倒産件数, 被害者, 商工リサーチ, 認証取得, 破産件数	可能性	上昇
販売競争, 激化, 売上高, 経常利益, 円, 税引き利益, ３月期, %, %減, サンウエブ	販売競争	上昇	店, 売り上げ, 出店, 店舗, 売上高, 減少, 大型店, 商店街, %, スーパー	売り上げ	下降
株価, 株, 銘柄, プット, 円, オプション料, 急落, 下落, 投資家, 買い方	株価	下降	生保, 日産生命, 契約者, 可能性, 受け皿会社, 予定利率, 保険, 解約, 劣後ローン, 保険料	可能性	上昇
下落率, 住宅地, 商業地, 地価, 下落, 東京圏, 傾向, 平均路線価, 前年, 大阪圏	傾向	発生	担保不動産, 回収, 下落, 土地, 地価下落, 住専, 価格, 価値, 目減り, 金融機関	価格	下降
景気, 兆し, 先行指数, 回復, 景気回復, 一致指数, 指標, 現状, 認識, プラス	兆し	発生	公益法人, 法人, 献金, 指定法人, 政治団体, 天下り理事, 日清医療食品, 実態, 補助金, 村田氏	実態	上昇
採算, %減, 悪化, 経常利益, 売上高, 円, 減少, 営業利益, %, 中間期	採算	下降	比率, %, 経常利益, 利益率, 売上高, 円, 営業利益, %増, 見通し, 採算	比率	上昇
声, 社長, 株主, 責任, 会長, 若狭氏, 退任, 辞任, 総会, 社内	声	上昇	インターネット, ホームページ, 動き, 情報, プロバイダー, 回線混雑, 茶道部, 有害情報, 情報提供, ボランティアネット	動き	発生
株価, アジア, 収益, 利益, 有利発行, 暴落, 追加負担, ドル高, 悪化, 低下	収益	下降	大蔵省, 破たん, 経営, 業務停止命令, 阪和銀行, 金融機関, 早期是正措置, 命令, 悪化, 異議申し立て	経営	下降
結論, ＮＡＯＣ, 議論, 公安審, 意見, ＦＩＳ, 検討, 協議, 運輸省, 問題	結論	発生	重油, 漂着, 油塊, 可能性, 油, 管区海上保安本部, 船体, 青森県沿岸, ナホトカ号, 重油流出事故	可能性	発生

表 10: 因果関係があると判断された知識 (b)

原因			結果		
ラベル	NP	属性	ラベル	NP	属性
流れ, 加速, 資金, 日本, 保, 円安, 民主党, の, こと, 選挙	流れ	上昇	首相, 可能性, 加藤氏, フン, 保, カンボジア, セン, 留任, 会談, 臨時国会	可能性	上昇
株価, 上昇, 市場, 急騰, 円, 株, 高値, ドル, 買い, 見方	株価	上昇	構想, 浮上, 社民党, 政府, 自民党内, 空中権, 与党内, 党, 与党, 酒造会社	構想	上昇
バブル, 人間, の, 自分, 心, 物, クルマ, それ, 火, 世代	バブル	上昇	考え方, 考え, 意向, の, 企業, %, こと, 日本, よう, 必要	考え	上昇
神戸, 震災, 思い, 優勝セール準備, 被災者, 復興, 被災地, 地元胴上げ, 仮設住宅, 街	思い	上昇	緊張感, 福留, 安田, 西武, 巨人, 練習, ヤクルト, チーム, シリーズ, イチロー	緊張感	上昇
小沢氏, 新進党, 確執, 羽田グループ, 党, 党内, 与党側, 改革, 細川氏, 不協和音	確執	上昇	発言力, 発言, 申告漏れ, 党, 日本, 中国, 自民党, 持永氏, の, 新進党	発言力	上昇
日切れ法案, 新進党, 成立, 審議, 影響, 参院, 衆院, 法案, 年度内, 暫定予算案	影響	発生	マイナス, 影響, 悪化, ポイント, %, 企業, 製造業, 消費税率引き上げ, —九月期, 東洋製鋼	影響	発生
イメージ, 龍, 酒, の, 伊達, 黒, 街, カービングスキー, マジック, 都甲	イメージ	上昇	景気, 金利, 回復, 設備投資, 上昇, 金利上昇懸念, 倒産件数, 水野清本部長ら, 公定歩合, 4 月	金利	上昇

表 11: 因果関係がないと判断された知識

原因			結果		
ラベル	NP	属性	ラベル	NP	属性
事件, 弁護人, 証人, の, 問題, サリン, 日本, よう, 何, 山一	事件	発生	拍手, 浅利選手, 位, 艇, 声援, 山尾, ボール, 真木選手, ゴール後, 組	拍手	上昇
人気, 水草, 一, 円, 河童, シーキングザパール, ラフティング, 魚, 水, キロ	人気	上昇	開示, 都, 混乱, 誤解, 接待相手, 無用, 相手名, 主張, 佐藤久夫裁判長, 開示決定	混乱	発生
私, 花, 先生, 心, 生徒, 旧暦, 春, 帽子, 年生, 話	花	上昇	可能性, 金大中氏, 議席, 投票, 与党, 決選投票, 投票率, 候補, 野党, 政権	可能性	上昇
火災, 発生, 火事, 商店街, 事故, 半焼, 台, 燃料タンク, 防火対策, 訓練	火災	発生	大統領, 可能性, 特赦, 大統領経験者, フジモリ大統領, 国家情報局, リマ, 赦免, 盧, 光州事件	可能性	上昇
パソコン, ワープロ, 機能, 野球道, ウィンドウズ, 野球ゲーム, 文書, ＰＰＲＡＭ, アウトレット品, ＭＰＥＧ	機能	上昇	地震, 群発地震, 活断層, 可能性, 断層, 牛伏寺断層, 地震調査研究推進本部, 活動, 地震活動, 地震調査委員会	可能性	上昇
県, 影響, 干潟, 住民, 工事再開, 市, 調査, 町長, 説明会, 反対	影響	発生	体, 仰木監督, 西武, 宙, オリックス, イチロー, 野村監督, 球場, ヤクルト, 東尾監督	体	上昇
依存度, 社, 大手, 山一, 株式委託手数料, 中小証券, 証券会社, 株式売買委託手数料, 半導体部門, 自由化	依存度	上昇	戸, 給水制限, 影響, 水, 取水制限, 利根川水系, 減圧給水, %, 出, 貯水率	影響	発生
症状, Ｏ１５７, 下痢, 検出, 血便, 菌, 女児, 感染, 腹痛, 食中毒	症状	発生	交流, 日本, 交渉, アジア, 将棋, 台湾, 人, ロシア, 大学, 中国	交流	上昇
サイン, バント, 西武, 回, 松井, 高木大, エンドラン, 巨人, 千住, 盗塁	サイン	発生	人質, 死傷者, ペルー政府, 大使, 青木大使, フジモリ大統領, トンネル, 強行突入, ゲリラ, ペルー	死傷者	発生
案, 浮上, 有力, 政府, 年度, 与党, 新進党, 検討, 公共事業, 方針	案	上昇	がん, 薬, 心筋梗塞, 遺伝子, 乳がん, 発病, 危険因子, 危険性, 遺伝的多様性, 患者	危険性	上昇
体, 仰木監督, 西武, 宙, オリックス, イチロー, 野村監督, 球場, ヤクルト, 東尾監督	体	上昇	件, 可能性, 帝国データ, 着服, 倒産, 倒産件数, 被害者, 商工リサーチ, 認証取得, 破産件数	可能性	上昇
支障, 企業, 野村, 信組, 社, 業務, 会長, 東証, 証券会社, 売買	支障	発生	成果, 練習, 原田, 選手, 関西大会, 競技, チームてこ入れ策, 優勝, チーム, 監督	成果	上昇

表 12: 過半数が選択した項目がなかった知識

原因			結果		
ラベル	NP	属性	ラベル	NP	属性
私, 気, 犬ベル, 人ごと, 急, とよのさん, そば, 施政方針演説, 気配, 学校	気	上昇	新聞, 情報, 電子メディア, 物知り, メディア, 国語力, 満載, 報道, 膨大, 記事	情報	上昇
検診, がん, 前立腺がん, 患者, 神経芽腫, 症状, 肺がん検診, 有効性, 生存率, 早期がん	症状	発生	動き, 下落, 取引, 上昇, %, 投資家, 野村, 円, 市場, 九月物	動き	発生
意見, 委員, 大勢, 公共事業, 中医協, 導入, 住民, 国, アセス, 結論	意見	上昇	女性, 関心, 葬式, いびき, 男性, 暴力, ブラックパール, ギャベ, 色, たるみ	関心	上昇
円安, ドル高, 声, 円相場, ドル, 円, 懸念, 日本, 市場, 円台	声	上昇	女性, %, 男性, 年齢, 独身志向, 歳以上, 人, 代, 就業継続, 結婚	年齢	上昇
小沢氏, 細川氏, 新進党, 違い, 党首選, 解釈, 羽田氏, 新規立法, 表面化, 鮮明	違い	上昇	大蔵省, 破たん, 経営, 業務停止命令, 阪和銀行, 金融機関, 早期是正措置, 命令, 悪化, 異議申し立て	経営	下降
支障, 企業, 野村, 信組, 社, 業務, 会長, 東証, 証券会社, 売買	支障	発生	株価, 期待, 平均株価, 市場, 円, 益出し, 失望, 戸田建設株, 低位大型株, 自社株買い	期待	上昇

C　抽出された因果関係知識の時系列データ

本手法により，因果関係があると判定されたイベント対の時系列データの一部を図 32~37 以下に示す．なお，図 32~34 は VAR モデルの次数が 1 と推定されたイベント対，図 35~37 は次数が 8 と推定されたイベント対である．

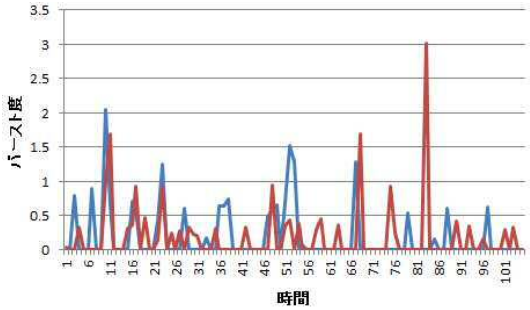

図 32: 次数 1 のイベント対 (a)

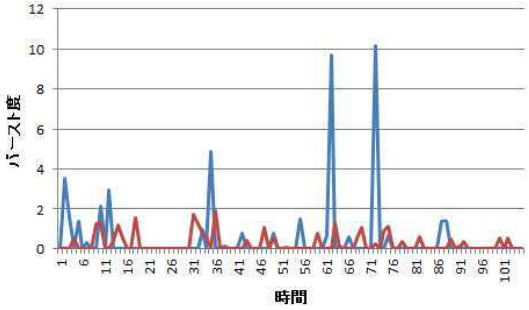

図 33: 次数 1 のイベント対 (b)

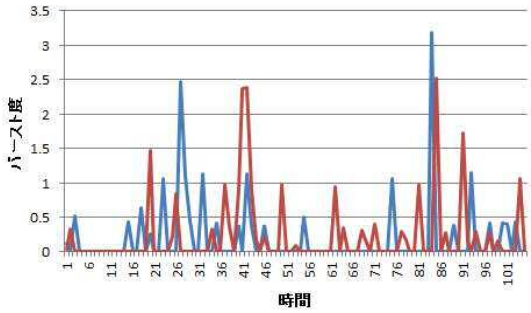

図 34: 次数 1 のイベント対 (c)

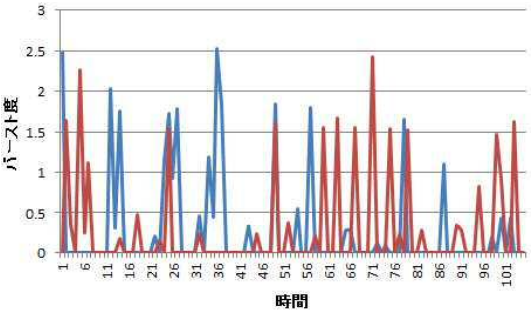

図 35: 次数 8 のイベント対 (a)

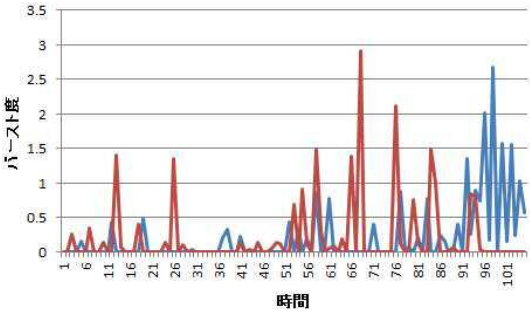

図 36: 次数 8 のイベント対 (b)

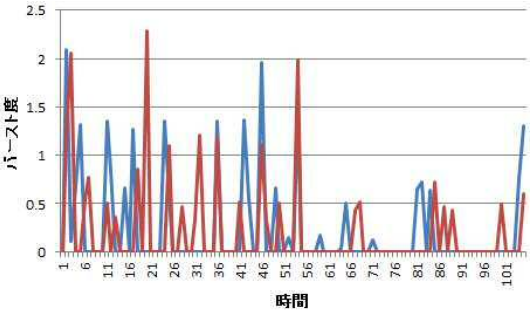

図 37: 次数 8 のイベント対 (c)